PRODUCTION AND OPERATIONS MANAGEMENT

POMS

Vol. 16, No. 1, January-February 2007, pp. 109–124
ISSN 1059-1478 | 07 | 1601 | 109$1.25

Pareto-Improving Contracts for Express Package Delivery Services

Candace A. Yano • Alexandra M. Newman

Department of Industrial Engineering and Operations Research and The Haas School of Business, University of California, Berkeley, California 94720, USA
Division of Economics and Business, Colorado School of Mines, Golden, Colorado 80401, USA
yano@ieor.berkeley.edu • newman@mines.edu

We address the problem of an express package delivery company in structuring a long-term customer contract whose terms may include prices that differ by day-of-week and by speed-of-service. The company traditionally offered speed-of-service pricing to its customers, but without day-of-week differentiation, resulting in customer demands with considerable day-of-week seasonality. The package delivery company hoped that using day-of-week and speed-of-service price differentiation for contract customers would induce these customers to adjust their demands to become counter-cyclical to the non-contract demand. Although this usually cannot be achieved by pricing alone, we devise an approach that utilizes day-of-week and speed-of-service pricing as an element of a Pareto-improving contract. The contract provides the lowest-cost arrangement for the package delivery company while ensuring that the customer is at least as well off as he would have been under the existing pricing structure. The contract pricing smoothes the package delivery company's demand and reduces peak requirements for transport capacity. The latter helps to decrease capital costs, which may allow a further price reduction for the customer. We formulate the pricing problem as a biconvex optimization model, and present a methodology for designing the contract and numerical examples that illustrate the achievable savings.

Key words: transportation contracts; contract pricing; speed-of-service pricing; time-of-use pricing; day-of-week pricing
Submissions and Acceptance: Received June 2005; revision received March 2006 and July 2006; accepted July 2006.

1. Introduction

Most package delivery companies (PDCs) charge a premium for faster delivery, but the practice of pricing by day of week is very limited. In the absence of this type of price differentiation, shipment volumes exhibit strong day-of-week patterns, especially in the express package delivery market. Although the schedules of various ground transport vehicles often can be adjusted to account for this day-of-week seasonality, express package delivery companies rely heavily on aircraft, for which it is not possible to match shipping capacity to demand very well. Consequently, excess shipping capacity varies by day of week and by route. When negotiating with potential high-volume contract customers, it may be advantageous to offer the customer an incentive to release packages counter-cyclically to the overall demand pattern. Such a counter-cyclical release pattern would improve the profit of the PDC in two ways. First, revenue is generated using available excess capacity for which the incremental operating costs are quite small. Second, by smoothing the overall demand pattern, requirements for additional transport capacity (typically provided by commercial carriers at premium prices) are minimized, and the PDC is able to provide more reliable service to all customers because the reduced peak loads pose less strain on pickup, delivery, and sortation resources. Because the incremental cost of servicing a contract customer with a counter-cyclical demand pattern may be small, the PDC may be able to pass on the savings to its customers by charging lower average prices.

Our research was motivated by a PDC whose management had hoped to induce the company's contract

customers to behave in the desired way via day-of-week and speed-of-service pricing alone. As we explain in more detail later, this is usually not possible. For this reason, we seek to develop a methodology for structuring contracts—which may include day-of-week and speed-of-service pricing as one element—that achieves the highest total profit for the PDC while ensuring that the customer is at least as well off as he would be under an existing contract or under any arbitrary reference price structure.

We examine this problem assuming that the PDC is negotiating with one major customer at a time. The most promising opportunities for improving the PDC's profit via more complex contract pricing arrangements occur in situations in which several customers sharing an aircraft route have similar day-of-week seasonality. This phenomenon occurs frequently due to weekly procurement cycles. For example, automobile assembly plants request deliveries of many parts on Monday morning to supply the assembly line for the week. (Although this may not be optimal, typical material requirements planning systems operate on a weekly schedule, and the procurement process follows suit.) Component suppliers in the same vicinity that provide parts to a given assembly plant therefore ship on the same day, usually Friday. The PDC would like all of these customers to modify their shipment plans, but it usually faces the problem of negotiating with them one at a time. When negotiating with a given customer, the PDC could consider likely outcomes of later negotiations with other customers, but this is obviously difficult to do because of the uncertainty involved. In our approach, various problem data can be specified to account for any capacity availability profiles (induced by non-contract customers and other contract customers) that the PDC wishes to consider.

In this paper, we focus on the flow of a class of homogeneous (or nearly homogeneous) packages from a single shipper (typically a manufacturer) that provides vendor-managed inventory (VMI) services to a single consignee (a downstream user of the manufactured parts). In the concluding section, we explain how our approach can be generalized to multiple package types. Because of the VMI arrangement, the shipper owns the goods and therefore incurs inventory holding costs until the consignee utilizes the goods. We emphasize that our approach is designed for situations in which the customer has considerable control over the timing of package releases which would usually entail changes in the production schedule, and thus our approach probably would not be suitable for an Internet retailer that is expected to fulfill orders soon after they arrive, often by a speed or mode of service chosen by the end-customer.

The remainder of this paper is organized as follows: The next section contains a review of the literature. This is followed by formal statements of the PDC's and customer's decision problems. In Section 4, we formulate the PDC's and customer's problems under a price-only contract and discuss the shortcomings of such a contract in our problem context, and this discussion provides a backdrop for our solution strategy. In Section 5, we present the details of our methodology for structuring Pareto-improving contracts. Section 6 provides numerical examples that illustrate our proposed method and its benefits. Section 7 closes the paper with a discussion of extensions and generalizations of our approach.

2. Literature Review

In this section, we provide an overview of the separate literatures on time-of-service pricing, and on speed-of-service and priority pricing. It is important to point out that, to the best of our knowledge, there is very little research that considers both simultaneously. We first discuss time-of-service pricing with an emphasis on electricity, toll roads, and computer and telecommunication network services, which are the most common application areas. Later in the section, we discuss the literature on speed-of-service and priority pricing, which tends to be less application-specific. In the interest of brevity, our citations are limited. Our intent is to provide the reader a sense of the issues that have been explored.

2.1. Time-of-Service Pricing

Vickrey (1971) provides a very lucid qualitative discussion of the benefits of what he calls "responsive pricing," that is, pricing that varies according to the state of the system. Responsive pricing includes such concepts as dynamic pricing based on instantaneous (real-time) congestion, time-of-service pricing based on forecasted (not real-time) demand or congestion patterns, and pricing schemes in the vein of current-day revenue management. Vickrey (1971) mentions application areas such as long-distance telephone service, airline reservations, and water and power delivery—the very same types of applications that motivate present-day research.

2.1.1. Electricity. Electricity markets are the most common application domain for time-of-use pricing, which is commonly referred to as peak load pricing in this industry. Here, peak prices have the effect of both reducing total demand and shifting some demand to off-peak periods. Most of the research can be classified into three broad areas: (1) the welfare economics of time-of-use pricing, (2) models of price elasticity for electricity, and (3) methods for setting prices. Seminal papers on the welfare benefits of peak-load pricing include Boiteux (1960) and Williamson (1966). Although much of the discussion is posed in terms of

peak versus off-peak prices, Panzar (1976) argues that capacity costs depend not only on peak loads but also on the loads during non-peak periods. Eckel (1987) examines the question of pricing based on demand-class (i.e., industrial, commercial, and residential consumers).

The literature on models of price elasticity for electricity is too extensive to discuss here. For a recent article, see Kamerschen and Porter (2004). These price elasticity models and estimates are widely used in pricing methods, where the emphasis is on setting prices during peak demand periods so as to attenuate demand and thereby reduce capacity requirements. Crew et al. (1995) provide a historical perspective on optimization-based time-of-use pricing approaches, focusing on non-storable goods such as electricity. Borenstein (2005) highlights several important issues in designing a pricing scheme for utility companies, including: (1) how often prices change, and (2) how long the delay between setting and realizing a price is. The most extreme, yet most effective, pricing scheme is real-time pricing. Borenstein describes various implementations of real-time pricing, and the implications of each. He notes that technology plays a key role in the effectiveness of real-time pricing.

2.1.2. Transportation. Although peak pricing is not yet widespread in transportation systems, researchers have been espousing the welfare gains and social benefits for years, citing the need to consider factors such as congestion externalities and environmental effects. See, e.g., an early paper by Vickrey (1963) and a more recent anthology edited by Button and Verhoef (1998). Wachs (2005) describes the current state of peak-load pricing on urban road networks, noting that only recently has technology enabled such pricing methods.

More recent research on time-of-day pricing for toll roads, bridges, tunnels, etc., has begun to consider the impact of traveler choices. Generally, these models assume that the traveler has the objective of minimizing some function of delay and out-of-pocket costs, and the toll setter chooses prices to maximize social welfare. Examples of papers in this stream of research include Arnott et al. (1990), Yan and Lam (1996), and Daganzo and Garcia (2000).

Two papers that treat models similar to ours are Brotcorne et al. (2000 and 2001). The authors address static problems in which the transport provider sets day-of-week (but not speed-of-service) prices and the customer chooses how much to ship on each day to satisfy some aggregate requirement over the horizon. An important simplifying assumption in these models is that the PDC has infinite capacity to handle each of the customer's shipping options.

2.1.3. Computer Network and Telecommunication Services. Computer network and Internet services represent another important application arena for peak-load pricing because of the very high amplitude of peaks that cause "busy signals" and slow transmission. At this writing, time-of-use pricing is rarely used, and many vendors promote flat rate, rather than usage-based, pricing. Researchers have modeled and demonstrated the benefits of pricing based both on usage alone and on usage in combination with induced congestion externalities (Gupta et al. 2001) for computer networks. For both computer network and Internet services, Paschalidis and Tsitsiklis (2000) suggest that time-of-day pricing alone, without adjustments for instantaneous congestion, may be sufficient to achieve good results for revenue and welfare maximization. In their model, customers are classified by average processing time per "call," and each customer class pays a different fee. In the context of pricing communications bandwidth, Altmann and Chu (2001) propose a combination of a flat fee that covers the cost of a basic level of bandwidth and usage-based charges for on-demand access to higher levels of bandwidth.

Interestingly, telecommunications service providers have long used time-of-day and day-of-week pricing, but as telecommunications capacity expands and competition becomes fiercer, vendors are offering more flat-rate, unlimited-use packages. These patterns are consistent with observations by Odlyzko (2001) who reports that for various communication technologies from regular mail to the Internet, as the technology matures, quality improves, prices fall, and pricing plans become simpler.

2.1.4. Time-of-Service Pricing in Other Industries. Increasingly more sophisticated time-of-service pricing—often called "revenue management" in recent years—has been adopted in industries in which many customers make purchases or reservations in advance. Gerstner (1986) examines peak-load pricing for private enterprises such as airlines, hotels, and restaurants. These scenarios differ from most of those above because of the need to consider competition, either directly or indirectly. For surveys, see Weatherford and Bodily (1992), Bitran and Caldentey (2003), and Talluri and van Ryzin (2004).

We now turn to a discussion of pricing based on speed of service or service priority.

2.2. Speed-of-Service or Priority Pricing

2.2.1. Queueing Models. The literature on congestion-dependent admission controls or pricing in queueing systems extends back several decades. Here, we provide a thumbnail sketch of the main problem categories and a few representative articles in each. Admission control involves deciding whether or not to admit customers to a system, considering the current congestion level. Most of the early work on ad-

mission control focuses on stationary systems with homogeneous customers. Examples of articles on this topic include Gavish and Schweitzer (1977) and Stidham (1985).

Researchers have also examined pricing as a means of admission control. When customers are homogeneous with respect to their valuation of the service and their disutility of waiting, researchers have considered two different ways to limit demand. One method uses static prices while the other uses congestion-dependent prices. Both methods have the effect of causing balking of customers who arrive when the queue is long, but the degree of control differs. Stidham (1992) and Chen and Frank (2001) are examples of papers in these categories. Other models employ a static (selected) price and a quoted leadtime as simultaneous controls on the net arrival process. Both static and congestion-dependent lead time quotes have been studied (e.g., Palaka et al. 1998, Webster 2002, and Plambeck 2004). Researchers have also analyzed systems in which a different price is charged for each priority class (e.g., Mendelson and Whang 1990; Rao and Peterson 1998) or for each class of customers, where customers are differentiated by how much they value the service and their service requirements (e.g., Ha 2001). Pricing has also been considered in competitive environments where priority or delivery time is one dimension on which firms compete for customers (e.g., Li and Lee 1994, and Chen and Wan 2003). All of these papers assume stationarity of the (gross) arrival process, and virtually all assume stationarity of the service capacity.

2.2.2. Approaches other than Queueing. A few other articles have considered speed-of-service pricing outside of a queueing framework. Several researchers have studied the problem of setting postal prices by speed of service. These models (e.g., Crew et al. 1990) have implicit stationarity assumptions (i.e., each time period has the same characteristics, and there is no carry-over inventory from one day to the next). Speed-of-service pricing also has been studied in the context of container shipping (see Holguin-Verasa and Jara-Diaz 1999), where more rapid delivery can be provided by positioning the container in the most accessible locations on a freighter and/or assigning the container to the most accessible dockside locations for pickup. The number of these "most accessible" locations is limited, and the focus is on pricing these scarce resources.

2.2.3. Pricing of Products or Services with Multiple Attributes. Wilson (1993) provides a systematic treatment of nonlinear pricing approaches. In one chapter of his book, he focuses on the pricing of products or services with multiple attributes or dimensions. He provides a thorough treatment of the problem under the assumption that the firm's costs and the customer's benefits are additively separable across quality attributes. Under these assumptions, the customer's utility and the firm's cost for any combination of product attributes can be collapsed to a scalar "score," which simplifies the analysis considerably. For this model, he presents an example of pricing Federal Express services when the two attributes are package weight and speed of service. He then presents approaches for priority pricing with an emphasis on service reliability, and for capacity pricing methods that include both capacity (i.e., load-dependent) and usage charges.

3. Problem Statement and Relationship to Literature

Recall that our research is motivated by the desire of a PDC to structure a long-term (e.g., annual) contract with a large customer who provides VMI services to its end-customer ("consignee"). Consequently, we envision scenarios with relatively stable package flow patterns from week to week. We model a situation in which the PDC's customer manufactures a single product (or family of similar products) and must arrange for transport of the goods to a consignee. The consignee's demands may vary by day of week, but the weekly pattern is assumed to repeat indefinitely. We assume that the customer ships all of the packages under consideration via the PDC with whom it is negotiating a contract. We have observed that it is quite common that manufacturers who ship large quantities utilize one transportation company for all non-exceptional shipments, partly for convenience and partly to take advantage of volume discounts that are implicit or explicit in the terms of the contract. The customer has manufacturing capacity constraints, and chooses how much to produce on each day of the week. Although these production quantities may vary by day of week, the weekly pattern is assumed to repeat indefinitely. Concurrently with his production decisions, the customer must decide how many packages to release each day by each speed of service to minimize his total cost of transporting goods and holding inventory while satisfying consignee demands on time. Because of the VMI arrangement, the customer pays for the (financial) cost of holding inventory at his own location, in-transit to the consignee (including any time at the PDC facility) and at the consignee.

The PDC has finite transport capacity. For ease of exposition, we assume there is only one mode and physical route on which the packages may travel on a given day, for example, on an air freighter flying from an airport near the manufacturer to an airport near the consignee. We assume that the schedule of transport

vehicles is fixed and that the available (internal) transportation capacity on each day is known (although our analysis can be used to aid in setting appropriate capacity levels). Any shipment quantity exceeding the internal capacity is subcontracted to a third-party transporter (e.g., a commercial airline) at premium rates. In practice, variable transportation costs and handling costs may differ little or not at all by day of week, but there are exceptions. For example, the PDC may ship packages by air freighter between San Francisco and Denver on Monday through Thursday, but may schedule trucks on Friday, because the trucks are less expensive to operate and the packages transported by truck will arrive in time for Monday morning deliveries. Incremental handling costs (at both the origin and destination) may be higher on some days of the week if crews are required to work overtime to accommodate the workload. We omit the variable handling and transportation costs for simplicity, but it is easy to include them, and their inclusion does not affect our solution strategy.

Although the financial costs incurred by the PDC for holding the packages are negligible because the inventory is owned by the customer (under the VMI arrangement), storage space is very limited and the PDC is unaccustomed to holding much, if anything, for more than the time required for sortation, so we impose a holding cost at the PDC to deter unnecessary storage. (In essence, this cost ensures that packages are sent on the earliest available vehicle, considering the priorities of the packages.) Thus, while a package is at the PDC, the PDC incurs the physical storage costs, but the customer continues to incur the financial holding costs.

Throughout our analysis, we assume that the PDC knows or can estimate all relevant customer data. Although this is not strictly true, our assumption is more realistic in this context than might be expected. For example, when customers use express package delivery services, insurance is usually included, and the customers are expected to declare a value for each package. Thus, the PDC need only estimate the customer's (annual) inventory holding cost rate to arrive at an estimate of the customer's unit holding cost. Similarly, in the context of negotiations, the customer is usually asked to provide information on shipping volumes and anticipated shipping patterns. For existing customers, information can be inferred from current behavior: In the absence of day-of-week price differentials, many customers release goods to the PDC as they are produced. Information on consignee demand patterns also may be obtained in the negotiation process.

Our problem characteristics differ from those considered in the literature in several important respects. First, in nearly all of the application domains discussed in the previous section, total demand is sensitive to prices. On the other hand, in our model, the total demand remains constant. Prices can be used to shift the demands from one period to another, but they do not attenuate the total demand. Both the feasibility and cost of a demand shift depend upon the timing and magnitude of the shift. Second, nearly all of the papers in the literature are concerned with settings in which service must be provided contemporaneously, and unsatisfied demand is lost. In our problem context, the PDC sometimes has flexibility regarding the time at which service is provided. For example, if a customer requires delivery within two days, the PDC has the option of sending the packages on a plane scheduled either tonight or tomorrow night. Third, our problem framework allows prices to be specified by both time of service (in this case, the pick-up date) and speed of service, and to the best of our knowledge, none of the articles in the literature treats this case. Furthermore, unlike models that use a queueing framework, our scenario requires a 100% guarantee of on-time service. Fourth, the standard approach of collapsing a multi-attribute pricing problem into one with a single attribute by mapping the multiple attributes to a scalar "score" cannot be used here because of the complications caused by differential congestion on different days of the week and the fact that service need not be provided contemporaneously. Finally, the two papers by Brotcorne et al. (2000 and 2001) that address a problem whose structure is similar to the "price-only" version of our problem (presented in the next section) are based on the assumption of infinite transportation capacity. Consequently, their models do not require shipping decisions on the part of the PDC after the customer has made his release decisions. Moreover, although their models can accommodate different speeds of service, they are designed for setting time-invariant tariffs, not for prices that may vary by day of the week.

We next present the PDC's and customer's decision problems under a "price-only" contract that permits price differentiation by day of week and speed of service, but with no other constraints or provisions imposed on the customer. An understanding of the price-only framework is essential for observing its shortcomings in our problem context.

4. The Price-Only Contract

The PDC faces two types of decisions: choosing a price for each day-of-week and speed-of-service pair, then choosing how to ship the resultant customer releases so that they arrive on time. The PDC's objective is to maximize profit subject to the requirement of delivering the customer's releases on time and operating within transportation capacity constraints (or paying

for any overflow transportation by a third party). Because we are implicitly assuming that the PDC is a monopoly, we impose upper and lower bounds on the total price of the contract with the customer. These bounds define the target range that the PDC believes will be competitive vis-a-vis the competition. Also, practical considerations require that on each day, the price for faster service should be higher (or no less) than that for slower service, and that for any given due date, the price should be non-decreasing as the release date approaches the due date. We refer to these constraints as (price) monotonicity constraints.

For any price schedule specified by the PDC, the customer must decide how much to produce and how much to release each day by each speed of service to satisfy consignee demands, taking into account the availability of goods determined by his own capacity-constrained production schedule. The customer incurs variable shipping costs, which depend on the prices selected by the PDC, as well as inventory holding costs for goods that are produced earlier than they are needed at the consignee, and wishes to minimize the sum of these costs. We assume that the cost of holding inventory at the customer (origin) is less expensive than the cost incurred for inventory either in transit to or at the consignee. This is consistent with the usual pattern of suppliers under VMI arrangements to avoid shipping too much inventory too early in order to limit the amount of goods on consignment at the consignee.

Before continuing, it is useful to point out that determining the best pricing structure is a three-stage game with two players. In the first stage, the PDC chooses a price structure for the contract with the goal of maximizing his profit, recognizing that the customer will optimize his releases in response to the contract, and that the PDC will be required to satisfy those demands and can choose to do so at minimum cost. In the second stage, the customer optimizes his releases to minimize total cost while satisfying consignee demands. In the third stage, the PDC satisfies customer requirements while minimizing total cost. Thus, we have a three-stage game in which one player (the PDC) makes the first and third sets of decisions and the other player (the customer) makes the second set of decisions. Games with such a structure have received essentially no attention in the research literature, and as we explain in more detail in Section 4.1, are extremely difficult when the second set of decisions involves a non-smooth response to the first set of decisions. For brevity, we present each player's decisions in a single optimization model, but it is useful to keep the three-stage game in mind as the discussion proceeds.

We now formulate the problems faced by the PDC and the customer under a price-only (linear price) arrangement. Note that all data are expressed in terms of a standard package, which we call a "unit." We use the terms "unit" and "package" (meaning standard package) interchangeably throughout the paper.

We assume a linear price structure because it would be too difficult for both the PDC and its contract customers to accommodate a nonlinear pricing schedule. One important reason is that the PDC currently prices only by speed of service, and very occasionally, by day of week, and by characteristics of the package itself (weight, volume, origin, destination, etc.), but not by total number of packages of each type or speed-of-service released by the customer on each day. Thus, the billing system is not set up to handle prices that are nonlinear in the package flows. Moreover, if a nonlinear pricing scheme were implemented, the customers might expect volume discounts (concave pricing structures) whereas the PDC would sometimes opt for convex pricing as a deterrent to large shipments on peak days. The PDC management perceived that convex pricing schemes would not be well-received by its customers and would likely lead to a loss of business.

Notation

Data:

h_t = holding cost per unit at the PDC on day t
h_t^o = holding cost per unit at the customer (origin) on day t
h_t^d = holding cost per unit at, or in transit to, the consignee (destination) on day t
M = variable transportation cost for (overflow) packages handled by a third-party transporter
R_t = production capacity at the customer on day t
D_t = demand at the consignee on day t
C_t = PDC's shipping capacity on day t
P_{LB}, P_{UB} = lower and upper bounds on total contract price, respectively
T = length of the time horizon

Decision variables:

For package delivery company:

$p_{t'd}$ = price per unit released by the customer on day t' with due date d
$y_{t'td}$ = number of units transported by the PDC on day t, having been released by the customer on day t' with due date d
$H_{t'td}$ = number of units held at the PDC on day t that were released by the customer on day t' with due date d
z_t = transportation overflow on day t (units)

For customer:

$x_{t'd}$ = number of units released on day t' with due date d
Q_t = number of units produced on day t
I_t = inventory at end of day t at the customer

The PDC's problem is:

$$(P) \quad \max \sum_{t'} \sum_{d} p_{t'd} x^*_{t'd} - \sum_{t'} \sum_{t} \sum_{d} h_t H_{t'td} - \sum_{t} M z_t$$

$$s.t. \quad \sum_{t'} \sum_{d} y_{t'td} \le C_t + z_t \quad \forall \ t \tag{1}$$

$$\sum_{t} y_{t'td} = x^*_{t'd} \quad \forall \ t', d \tag{2}$$

$$H_{t't'd} = x^*_{t'd} - y_{t't'd} \quad \forall \ t', d \tag{3}$$

$$H_{t'td} = H_{t',t-1,d} - y_{t'td} \quad \forall \ t', t > t', d \tag{4}$$

$$p_{t'd} \ge p_{t',d+1} \quad \forall \ t', d \tag{5}$$

$$p_{t'd} \le p_{t'+1,d} \quad \forall \ t', d \tag{6}$$

$$P_{LB} \le \sum_{t'} \sum_{d} p_{t'd} x^*_{t'd} \le P_{UB} \tag{7}$$

$$p_{t'd} \ge 0 \quad \forall \ t', d \tag{8}$$

$$y_{t'td} \ge 0 \quad \forall \ t', t, d \tag{9}$$

$$H_{t'td} \ge 0 \quad \forall \ t', t, d \tag{10}$$

$$z_t \ge 0 \quad \forall \ t \tag{11}$$

where $x^*_{t'd}$, $t' = 1, \ldots, T$, $d \ge t'$ is the optimal release schedule from the customer's problem:

$$(C) \quad \min \sum_{t'} \sum_{d} p^*_{t'd} x_{t'd} + \sum_{t'} h^o_{t'} I_{t'} + \sum_{t'} \sum_{d} (d - t') h^d_{t'} x_{t'd}$$

$$s.t. \quad I_{t'} = I_{t'-1} + Q_{t'} - \sum_{d} x_{t'd} \quad \forall \ t' \tag{12}$$

$$0 \le Q_{t'} \le R_{t'} \quad \forall \ t' \tag{13}$$

$$\sum_{t'} x_{t'd} = D_d \quad \forall \ d \tag{14}$$

$$I_{t'} \ge 0 \quad \forall \ t' \tag{15}$$

$$x_{t'd} \ge 0 \quad \forall \ t', d \tag{16}$$

where $p^*_{t'd}$, $t' = 1, \ldots, T$, $d \ge t'$ are the optimal prices from the PDC's solution. Note that we have represented finite horizon versions of both problems, but the repeating weekly problem can be represented similarly with appropriate modulo arithmetic for the time indices. We have also assumed instantaneous PDC shipping time as reflected in constraints (3)–(4), but this can be adjusted by including the appropriate lag in the time indices.

The terms in the PDC's objective function are, respectively, revenue from the packages shipped by the customer, storage costs and the cost of overflow transportation. Constraint set (1) ensures that daily shipments do not exceed transport capacity or that overflow transportation is utilized at a cost, and constraint set (2) ensures that packages are shipped in time to satisfy the speed of service requested by the customer. Inventory balance constraints at the PDC appear in (3) and (4). Constraint set (5) ensures monotonicity of prices by speed of service on each day, and constraint set (6) ensures that for the same due date, prices are monotonically non-decreasing in the release date. Upper and lower bounds on the total contract price are reflected in (7). Finally, non-negativity constraints on the prices, shipment quantities, inventory, and overflow shipments are represented in (8)–(11).

The customer seeks to minimize total transportation and inventory holding costs. Because of the VMI arrangement, the customer incurs the cost of inventory held at his own location, as well as holding costs for all released goods until the respective due dates at the customer. Thus, any release of goods on day t' and due on day d is held for $d - t'$ days, and these costs are reflected in the third term of the objective. Inventory balance constraints at the origin are shown in constraint set (12). Constraints (13) ensure that production does not exceed capacity, while constraints (14) ensure that the customer releases goods and specifies due dates in such a way that the consignee's demands are satisfied on time. Note that constraints (14) are based on the assumption that the customer will not release packages earlier than would be required by the slowest speed of service. This is justified under a VMI arrangement because VMI suppliers prefer not to have excess inventory on consignment at the consignee, but may be willing to ship a few days early to save on transportation costs. (If one wishes to allow the customer to ship very early, this can be accommodated by expanding the set of $\mathbf{x}$ variables to include variables corresponding to earlier release dates for a given due date.) Constraints (15) restrict inventory to remain non-negative, which guarantees that goods are not released before they are available. Finally, constraint set (16) represents non-negativity of customer release quantities.

We note that if the customer is concerned about period-to-period variability of the production quantity, it is possible to include piecewise linear, convex production costs. The essential structure of the customer's problem remains the same, i.e., it remains a linear program when the PDC's prices are fixed. We expect that the inclusion of convex production costs will increase the difficulty of solving the overall problem because these costs implicitly limit the customer's flexibility, but technically, there is nothing that precludes such cost structures.

4.1. Solution Challenges

An ideal outcome in our problem scenario would be a solution in which the total system-wide cost is minimized. With such an outcome, the potential total savings is maximized, and the only remaining decision is how that savings should be shared. The system-wide

cost can be minimized by solving the centralized problem-choosing customer release and PDC shipment quantities to minimize total cost subject to PDC transportation capacity constraints and customer production capacity constraints. Unfortunately, there may be many alternate optima for the centralized problem, and it is difficult to design a price vector without knowing which of the optima is the solution that one wants to induce the two parties to choose voluntarily, and whether there even exists a monotonic price vector that will induce those choices. Indeed, in preliminary tests, we found that monotonicity constraints are especially problematic to satisfy, not only in this context but in others described later in this section.

For fixed prices, the customer's problem is a transportation problem, with the sources being the production capacity on each day and the sinks being the consignee demands on each day. The PDC's problem is a linear program for a fixed customer release plan (specifying quantities by speed of service on each day). Consequently, each problem is quite easy to solve if the other party's decisions are known, but because they are not known, the two sets of decisions are inextricably intertwined. In particular, for the PDC, the "other party's decisions" are the customer's optimal releases in response to the selected prices.

Observe that because of the "price only" (linear price) nature of the contract, there is no assurance that that the PDC is able to send the customer's shipments without the use of overflow capacity at a high cost. In addition, the solution of the customer's problem is an extreme point of his feasible region. Therefore, small changes in the prices may lead the customer to choose another extreme point that is qualitatively quite different, making it more difficult to use price setting as a means to manage capacity utilization than in situations where the customer's responses to price changes are smooth. Further, except for rare instances in which it is possible for the PDC to transport all goods on the same day they are released by the customer, one cannot check for feasibility of the customer's shipping schedule with respect to the PDC's transportation capacity constraints by inspection, but must solve the PDC's package flow problem to make this assessment. Indeed, this is the essential difficulty of the problem, and is the primary feature that differentiates it from existing models for transportation pricing.

In addition to the solution strategy described in the next section, we explored several other approaches which we describe briefly for the benefit of interested researchers. A few approaches were based on the centralized problem in which a single decision-maker seeks to unilaterally optimize all decisions. Using such a scheme, it is easy to minimize the total cost of the system. We explored whether the dual solution of the centralized problem could be used as the basis for constructing a price schedule. We found that the dual prices were not useful for this purpose because too many of the dual prices were equal to zero or equal to each other. Moreover, any selected linear price schedule has the same deficiencies as those described earlier. We also considered and explored using dual price information to construct piecewise linear convex pricing schemes, but as mentioned earlier, these were regarded as unrealistic by the PDC that motivated our research, and we found that it was difficult to incorporate workable monotonicity constraints with such a complex pricing framework. Finally, the solutions based on the centralized problem tended to be undesirable for the customer, not only from the standpoint of cost, but also from the standpoint of solution structure (e.g., variability of production, quantity of inventory held, etc.).

We also explored inverse optimization methods (see Ahuja and Orlin 2001), seeking to find the price vector closest to some initial price vector that would lead to a customer-selected release plan that was also feasible for the PDC. We found that it was difficult to incorporate monotonicity constraints on the prices and still achieve useful results. In virtually all cases, price monotonicity constraints made the problems infeasible.

Finally, we developed and tested a number of iterative methods designed to progressively adjust solutions toward those with lower cost, but these methods tended to terminate in price schedules that lacked monotonicity or customer releases that necessitated considerable overflow transportation.

Thus, the approach in the next section is based on the observations described above that: (1) price-only schedules will rarely produce solutions that are acceptable for the PDC because of the extreme point nature of the customer's releases under linear price schedules, (2) nonlinear pricing schemes are not only unpalatable to the PDC but also difficult to construct, and (3) methods that do not explicitly account for price monotonicity are unlikely to produce reasonable prices. Two other key factors in the design of our approach are the recognition that the customer's perspective and costs need to be considered, and that the main purpose of any optimization-based tool is to aid decision-making (or, in this case, to facilitate negotiations), not necessarily to specify a single solution.

5. Solution Strategy

Our proposed methodology for structuring a contract recognizes that there is almost always a reference price structure that serves as a starting point. For example, it may be the current price schedule offered to the customer. The price structure may include dif-

ferentiation by speed of service and/or day of week. At the PDC that motivated our research, many contracts have "flat" pricing with no differentiation, and customers release packages with the expectation of delivery via the fastest available speed of service on the given route.

We wish to structure a contract that maximizes the PDC's overall profit while ensuring the customer is at least as well off as he was under the reference price schedule **p**. This presents the opportunity for a win–win situation. We assume that the customer has solved (**C**) under the reference prices; we refer to the resulting solution as $\hat{\mathbf{x}}$. To determine the PDC's profit under the reference prices and given the customer's selected $\hat{\mathbf{x}}$, the PDC solves (**P**) with the values of **p** and $\mathbf{x} = \hat{\mathbf{x}}$ fixed, and without constraints (5)–(8), which are not relevant once the prices are fixed. We refer to the solution of this problem as $\hat{\mathbf{y}}$. Thus, **p**, $\hat{\mathbf{x}}$, and $\hat{\mathbf{y}}$ define the initial (benchmark) solution.

The PDC seeks to identify a contract that consists of customer releases, **x**, and per-unit discounts, $\delta_{t'd}$, where the net prices are $p_{t'd} - \delta_{t'd}$. The discounts serve to compensate the customer for any costly deviations of **x** from $\hat{\mathbf{x}}$. We impose a "generous" form of the constraints ensuring the customer is at least as well off as he was under the reference prices. In particular, we ensure that the total cost incurred by the customer to transport each day's demand is no greater than it was under the reference pricing schedule and the customer's original solution $\hat{\mathbf{x}}$. As an alternative, one could impose a single constraint ensuring that the customer's total cost under the new contract is no more than it was under the pre-existing terms. As before, we impose monotonicity constraints on the net prices.

In determining the contract parameters, the PDC must consider the customer's production capacity constraints and its own transportation capacity constraints—or the cost of overflow transportation if the constraints are violated.

The PDC's contract structuring problem is:

$$(\mathbf{P}_\delta) \quad \max \sum_{t'} \sum_{d} (p_{t'd} - \delta_{t'd}) x_{t'd} - \sum_{t'} \sum_{t} \sum_{d} h_t H_{t'td} - \sum_{t} M z_t$$

$$\text{s.t.} \qquad (1)\text{–}(4) \text{ and } (9)\text{–}(16)$$

$$\sum_{t} (h_t^d - h_t^o) \left(\sum_{t' \le t} x_{t'd} - \sum_{t' \le t} \hat{x}_{t'd} \right) \le \sum_{t'} p_{t'd} \hat{x}_{t'd} - \sum_{t'} (p_{t'd} - \delta_{t'd}) x_{t'd} \quad \forall\ d \qquad (17)$$

$$p_{t'd} - \delta_{t'd} \ge p_{t',d+1} - \delta_{t',d+1} \quad \forall\ t', d \qquad (18)$$

$$p_{t'd} - \delta_{t'd} \le p_{t'+1,d} - \delta_{t'+1,d} \quad \forall\ t', d \qquad (19)$$

$$\delta_{t'd} \ge 0 \quad \forall\ t', d \qquad (20)$$

$$p_{t'd} - \delta_{t'd} \ge 0 \quad \forall\ t', d \qquad (21)$$

The first term in the objective function represents the total revenue considering the discounts provided to the customer with the new release schedule x. The other terms capture the PDC's inventory holding and transportation overflow costs.

Constraints (17) ensure that, for each due date, the discounts reduce the customer's cost so that it is less than or equal to his cost incurred by the initial solution $\hat{\mathbf{x}}$. This is accomplished by ensuring that the additional holding cost incurred by the customer from using **x** rather than $\hat{\mathbf{x}}$ (the left hand side of the constraint) is less than the cost reduction between the old release schedule and the new one (the right hand side of the constraint). Observe that $\sum_{t' \le t} x_{t'd}$ represents the cumulative releases through day t of packages due on day d under release schedule **x** and $\sum_{t' \le t} \hat{x}_{t'd}$ represents the cumulative releases through day t of packages due on day d under schedule $\hat{\mathbf{x}}$. The difference between these terms represents the additional unit-days of inventory (which may be negative) either in transit to, or at, the consignee on day t that are due on day d. The incremental cost per unit for holding inventory either in transit or at the consignee versus at the customer is $h_t^d - h_t^o$. Thus, by properly weighting each quantity of unit-days by its incremental cost, then summing over all t, we can express the total incremental inventory cost incurred by the customer from using **x** rather than $\hat{\mathbf{x}}$. Constraints (18) and (19) ensure monotonicity of net (discounted) prices. Constraints (20) ensure nonnegativity of the δ values, while constraints (21) ensure nonnegative prices.

We make a few observations about the optimal solutions for $(\mathbf{P}_\delta)$. First, because of the structure of the formulation, the solution is the PDC's most profitable alternative subject to the constraint that the customer is no worse off than under the solution of (C) with the reference prices. Consequently, the solution is (weakly) Pareto-improving relative to the solution under the reference prices, and is Pareto-optimal for the PDC. In general, the customer would be better off if he were allowed to optimize his releases under the selected discounts. (Recall that in our contract structuring problem, the PDC chooses **x** for the customer but compensates him for any additional costs incurred.) However, allowing the customer to do so would have negative repercussions for the PDC, as the customer-selected release schedule would be an extreme point and have its associated disadvantages. This is the reason why the contract must include specification of **x** as well as the prices.

$(\mathbf{P}_\delta)$ is a bilinear optimization problem. That is, the

decision variables can be partitioned into two sets such that the remaining problem is a linear program if one set of variables is fixed. Notice that our problem is a linear program in the δ values if the other variables are fixed, and linear in the other variables if the δ values are fixed. Our problem is a jointly constrained bilinear program because the constraints cannot be partitioned to decouple the δ values from the remaining variables. Al-Khayyal (1990) describes how to utilize his algorithm (Al-Khayyal and Falk 1983) for jointly constrained biconvex optimization problems to optimally solve jointly constrained bilinear programming problems. Al-Khayyal's algorithm was designed as research code and is not publicly available. However, the BARON (Sahinidis 2000) software for nonconvex (non-concave) optimization can be applied to our problem. Recognizing that convex (concave) optimization software is in more widespread use, we also applied a commercial nonlinear solver (MINOS) using many randomly-generated starting points with the goal of assessing the quality of the best of these solutions.

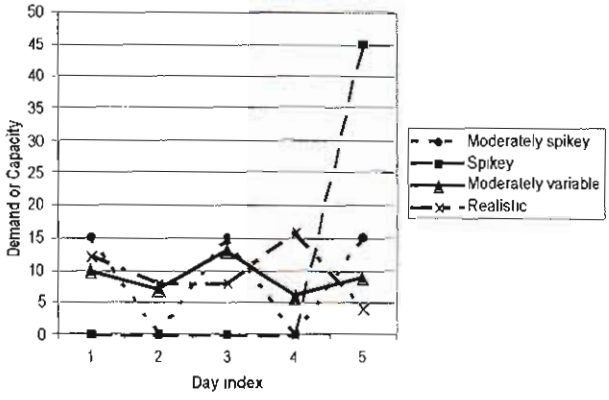
6. Numerical Results

We constructed a set of test problems with a view toward understanding how different patterns of demands, transportation capacities, production capacities, and initial prices would affect our ability to find good solutions. Table 1 lists the 32 combinations of demand patterns, transportation capacity patterns and production capacity patterns that we used. We generated twenty problem instances for each combination of patterns listed in Table 1, and examples of the various patterns are shown in Figure 1. All problems have a five-day weekly cycle. For the cases with constant demands, we set $D_t = 10$ for all t. For the moderately variable capacity patterns, we randomly generated values from a Uniform distribution on (0, 20) and rounded down to the nearest integer for all t. For the moderately spikey demand patterns, we generated quantities for days $t = 1, 3, 5$ from a Uniform distribution on (14, 18) and rounded each down to the nearest integer and for $t = 2, 4$, we set demands to zero. For the spikey case, demands or capacities were generated such that quantities for $t = 1, \ldots, 4$ were set equal to zero and the quantity for $t = 5$ was generated from a Uniform distribution on (44, 55) and rounded down to the nearest integer. We also generated demands according to a "realistic" scenario designed to mimic day-of-week shipments of a large company. To this end, we set demand on Monday to a number generated according to a Uniform distribution on the interval (10, 15) and rounded down to the nearest integer. We generated demands on Tuesday, Wednesday, Thursday, and Friday in a similar manner, but on the intervals (5, 10), (8, 13), (14, 19), and (3, 8), respectively. Having generated the demands, if the generated (total) production capacity level was insufficient to satisfy demand, we started the demand and capacity generation process over and repeated the process until a feasible pair of demand and production capacity vectors was generated. Similarly, if the generated (total) transportation capacity was less than 90% of the

Table 1 Scenario Attributes

Demand pattern	Transportation capacity pattern	Production capacity pattern
Constant	Moderately variable	Moderately variable
Constant	Moderately variable	Spikey
Constant	Spikey	Moderately variable
Spikey	Moderately variable	Moderately variable
Spikey	Moderately variable	Spikey
Spikey	Spikey	Moderately variable
Moderately spikey	Moderately variable	Moderately variable
Moderately spikey	Moderately variable	Spikey
Moderately spikey	Spikey	Moderately variable
Realistic	Moderately variable	Moderately variable
Realistic	Moderately variable	Spikey
Realistic	Spikey	Moderately variable
Constant	Constant	Moderately variable
Constant	Constant	Spikey
Spikey	Constant	Moderately variable
Spikey	Constant	Spikey
Moderately spikey	Constant	Spikey
Moderately spikey	Constant	Moderately variable
Realistic	Constant	Moderately variable
Realistic	Constant	Spikey
Constant	Moderately variable	Constant
Constant	Spikey	Constant
Spikey	Moderately variable	Constant
Spikey	Spikey	Constant
Moderately spikey	Moderately variable	Constant
Moderately spikey	Spikey	Constant
Realistic	Moderately variable	Constant
Realistic	Spikey	Constant
Constant	Constant	Constant
Spikey	Constant	Constant
Moderately spikey	Constant	Constant
Realistic	Constant	Constant

Figure 1 Representative demand and capacity patterns.



total demand, we eliminated the capacity vector and generated another. (Recall that transportation overflows are allowable at a cost, so total transportation capacity need not be greater than total demand.)

We set the holding cost to 0.02 per unit per period at the customer, 0.20 per unit per period at the consignee, and 1.00 per unit per period at the PDC. We set the cost at the consignee to be larger than the cost at the customer so that solutions would reflect usual behavior in a VMI arrangement. That is, the supplier (the customer in our framework) prefers not to have excess inventory on consignment at the consignee, and the higher holding costs at the consignee deter him from doing so. We set the holding costs at the PDC to be larger than the other two values to induce the usual PDC behavior of shipping as soon as possible.

We used two sets of reference prices: "flat" and speed-of-service. For the flat prices, we set $p_{t'd} = 40$ for all t' and d. Under flat pricing, the customer releases "just-in-time." For speed-of-service pricing, we set $p_{t',t'} = 40.00$, $p_{t',t'+1} = 39.00$, and $p_{t',t'+2} = 38.90$. Under this price schedule, and with the holding costs given earlier, the customer's economic tradeoffs (considering his transportation and holding costs) cause him to prefer the middle speed of service among the three available to him. Our intent here was to choose initial prices that would lead to solutions from (C) that are qualitatively different than the "just-in-time" releases under flat pricing. The speed-of-service pricing structure that we used is structurally similar to the price structure that is offered to non-contract customers by the PDC that motivated our research. For example, when the options are "same-day," "one-day," and "two-day" service, one-day service is substantially cheaper than same-day service, but the additional discount for two-day service is often quite modest. We set the cost of overflow transportation (per unit) to 50, reflecting that the cost of using a commercial carrier usually exceeds the revenue.

For each scenario, we first solve the customer's problem (C) with the references prices and note his release schedule, i.e., the $\hat{\mathbf{x}}$ values. We use these release values to solve for $\hat{\mathbf{y}}$ to provide a benchmark, i.e., the PDC's profit resulting from his optimal shipping schedule, given the customer's release schedule in reaction to the reference prices. Then, we solve the PDC's revised problem, ($\mathbf{P}_\delta$). We solve each instance of ($\mathbf{P}_\delta$) in two ways: (1) applying the BARON non-convex optimization software, and (2) applying a nonlinear (convex) solver (MINOS) to ($\mathbf{P}_\delta$) using 250 different starting points.

The BARON solver produces an ϵ-optimal solution (within a user-specified ϵ). We specified $\epsilon = 0.1$, which is a very small fraction of the typical profit levels (virtually all between 1000 and 2000) in our problem instances. BARON produced verified optimal solutions in the vast majority of the problems, and ϵ-optimal solutions in a handful of instances. The CPU times are a small fraction of a second.

For the second approach, we independently generated each initial $\delta_{t'd}$ from a continuous uniform distribution on the interval $[0, P^{max}]$, where P^{max} represents the highest price among the reference prices. CPU times are only a small fraction of a second for each of the 250 starting points. For each problem instance, we identify the best of the 250 solutions and compute the resulting profit. Using the BARON solution as the basis for comparison, we were able to confirm that the best of the solutions from 250 starting points using MINOS was either optimal or very near optimal (within pennies).

A more detailed study of the 250 solutions for the various problem instances led to surprising results. Not only was the best of the solutions either optimal or very close to optimal, but in only 44 of the 640 cases with a constant initial price vector, and in 20 of the 640 cases with a speed-of-service-only initial price vector was the worst of the solutions more than 1% from optimal. We investigated a large sample of these problem instances in more detail and discovered that in every case, 249 of the 250 solutions were either optimal or within 0.5% of optimality, and only one solution was far from optimal. These results suggest the following:

(i) In a very high percentage of the problem instances, the objective function is either well-behaved, or all of the local optima have very similar objective values.

(ii) In some problem instances, there may be more than one local optimum, but if so, nearly all of them have similar objective values.

(iii) Applying an ascent procedure using a few starting points is likely to produce very good, if not optimal, solutions.

Although the BARON solver is fast and accessible, convex optimization software packages are more widely used in practice. Our results suggest that using a non-convex optimization package may not be essential, thereby facilitating the implementation of our solution approach.

We use the best of the available solutions as the basis for calculating savings. In some problem instances, due to high transportation overflow costs, the PDC's profit may be negative under the reference prices and the customer's initial solution $\hat{\mathbf{x}}$. For this reason, we calculate savings as a percentage of the sum of holding and overflow costs from $\hat{\mathbf{y}}$, which is the maximum savings that the PDC can achieve (i.e., the PDC's controllable costs). For some problem instances, the initial solution (from $\hat{\mathbf{y}}$) had no controllable costs for the PDC. In these cases, there is no need

to modify the prices. For these instances, we report the percentage savings as zero.

We report the minimum, average, and maximum percentage savings by scenario (across the 20 problem instances for each scenario) in Table 2 for cases starting with flat pricing, and in Table 3 for cases starting with speed-of-service pricing.

For the 32 scenarios with flat initial prices, the average savings is 64.7% with a standard deviation of 41.4%, while for the scenarios with speed-of-service-based initial prices, the average savings is 64.3% with a standard deviation of 37.6%. (If the instances with zero controllable PDC costs in the initial solution were omitted, the average percentage savings would be higher and the standard deviation lower.) The high standard deviations are due to the fact that for some problems, it is impossible to achieve any savings by a rearrangement of the customer's release schedule, while for others, all controllable costs can be eliminated by doing so. In particular, for cases with 100% savings, the PDC is able to eliminate all of its storage costs and completely avoid the use of overflow shipments by using the solution from our approach.

Not surprisingly, savings are zero or negligible in most scenarios with spikey transportation capacity. In these scenarios, even extreme price changes may not induce the customer to release packages when transportation capacity is most available, because the customer has production capacity constraints that prevent him from doing so. Only when both transportation and production capacity are spikey (with spikes in the same period) will discounts from an initial speed-of-service price vector lead to sizable savings. For scenarios without spikey transportation capacity, the average savings ranged from 34.8% to 100% of the PDC's controllable costs. Of course, the degree to which savings can be achieved is data-dependent, but these results suggest that even when consignee demand fluctuates substantially and transport capacity is not well aligned with consignee de-

Table 2 Savings as Percent of Controllable Costs for 32 Scenarios Starting with Flat Pricing

(Demand, transportation capacity, production capacity)	Minimum savings (%)	Average savings (%)	Maximum savings (%)
(c, mv, mv)	80.5	97.8	100.0
(c, mv, s)	79.0	97.7	100.0
(c, s, mv)	0.0	0.0	0.0
(s, mv, mv)	31.6	55.4	100.0
(s, mv, s)	20.7	58.6	87.5
(s, s, mv)	0.0	0.0	0.0
(ms, mv, mv)	78.4	95.5	100.0
(ms, mv, s)	80.7	98.3	100.0
(ms, s, mv)	0.0	0.0	0.0
(r, mv, mv)	36.5	88.9	100.0
(r, mv, s)	86.0	98.4	100.0
(r, s, mv)	0.0	0.0	0.0
(c, c, mv)	77.2	96.3	100.0
(c, c, s)	44.0	92.9	100.0
(s, c, mv)	45.5	51.1	57.1
(s, c, s)	47.6	53.8	57.1
(ms, c, s)	100.0	100.0	100.0
(ms, c, mv)	100.0	100.0	100.0
(r, c, mv)	64.0	94.9	100.0
(r, c, s)	34.2	96.7	100.0
(c, mv, c)	52.9	92.3	100.0
(c, s, c)	0.0	0.0	0.0
(s, mv, c)	15.2	53.4	94.4
(s, s, c)	0.0	0.0	0.0
(ms, mv, c)	81.9	98.6	100.0
(ms, s, c)	0.0	0.0	0.0
(r, mv, c)	83.5	98.0	100.0
(r, s, c)	100.0	100.0	100.0
(c, c, c)	50.0	53.1	57.1
(s, c, c)	100.0	100.0	100.0
(ms, c, c)	100.0	100.0	100.0
(r, c, c)	100.0	100.0	100.0

c = constant, mv = moderately variable, s = spikey, ms = moderately spikey, r = realistic

Table 3 Savings as Percent of Controllable Costs for 32 Scenarios Starting with Speed-of-Service Pricing

(Demand, transportation capacity, production capacity)	Minimum savings (%)	Average savings (%)	Maximum savings (%)
(c, mv, mv)	39.8	92.2	100.0
(c, mv, s)	15.3	88.7	100.0
(c, s, mv)	1.0	1.0	1.0
(s, mv, mv)	7.8	40.1	100.0
(s, mv, s)	1.1	42.8	77.5
(s, s, mv)	31.0	88.0	100.0
(ms, mv, mv)	10.2	82.7	100.0
(ms, mv, s)	59.5	96.1	100.0
(ms, s, mv)	1.6	1.9	2.2
(r, mv, mv)	9.2	78.0	100.0
(r, mv, s)	41.3	95.1	100.0
(r, s, mv)	0.2	0.4	0.6
(c, c, mv)	21.3	86.7	100.0
(c, c, s)	9.5	84.1	100.0
(s, c, mv)	29.8	34.8	40.5
(s, c, s)	31.7	37.3	40.5
(ms, c, s)	100.0	100.0	100.0
(ms, c, mv)	100.0	100.0	100.0
(r, c, mv)	15.3	80.8	100.0
(r, c, s)	8.0	95.4	100.0
(c, mv, c)	6.5	79.8	100.0
(c, s, c)	1.0	1.0	1.0
(s, mv, c)	5.5	36.6	89.6
(s, s, c)	18.7	83.2	100.0
(ms, mv, c)	72.0	97.2	100.0
(ms, s, c)	1.6	1.9	2.2
(r, mv, c)	63.9	95.9	100.0
(r, s, c)	0.2	0.4	0.6
(c, c, c)	100.0	100.0	100.0
(s, c, c)	33.8	36.6	40.5
(ms, c, c)	100.0	100.0	100.0
(r, c, c)	100.0	100.0	100.0

c = constant, mv = moderately variable, s = spikey, ms = moderately spikey, r = realistic

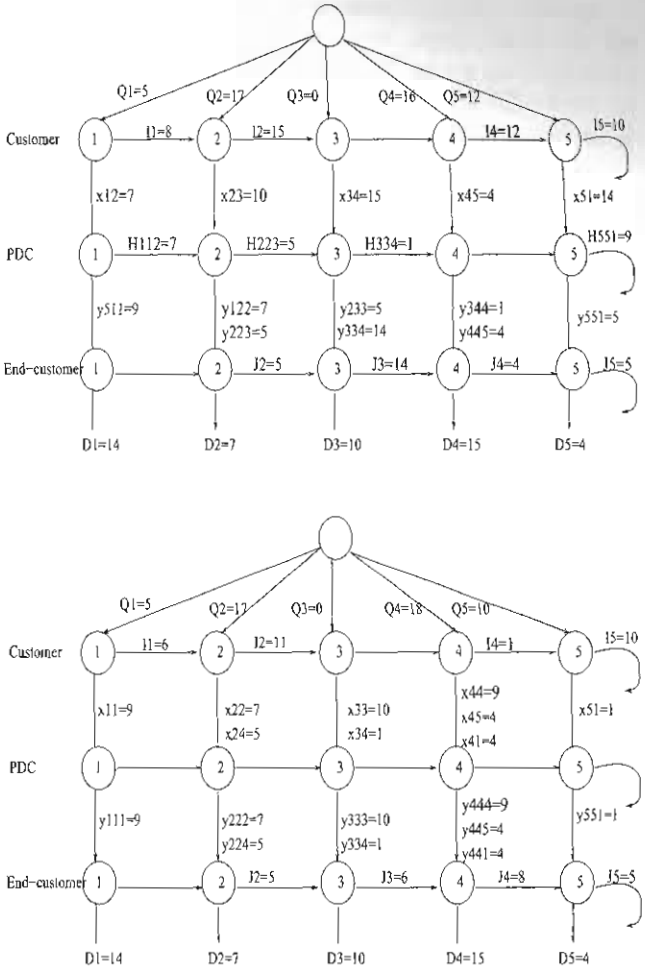
mand, the PDC can reduce its controllable costs by a substantial fraction while ensuring that the customer is no worse off, and in most realistic cases can be made strictly better off if the PDC passes on some of the savings to the customer.

We examine two problems in greater detail in order to illustrate how and why our solution procedure provides benefits. In these examples, the customer has the options of instantaneous (same-day), one-day, or two-day service. We first consider Example 1, in which the initial price vector reflects speed-of-service prices. Production capacities are moderately variable and are given by the vector (5, 17, 0, 18, 12), as are transportation capacities which are given by the vector (9, 12, 19, 17, 1). Demands are shown on the network diagram in the upper portion of Figure 2, along with production quantities and product flows under $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. (In Figures 2 and 3, J denotes physical inventory at the consignee.) With this pricing scheme, the customer has an incentive to release packages one day early. He produces on all days on which he has production capacity, and often produces ahead of schedule, not only because of production capacity constraints but also to take advantage of discounts for early releases. Given the customer's release schedule, the PDC ships the units to the consignee on or before their respective due dates. (On four of the five days, inventory is held at the consignee.) Although the customer's one-day-in-advance release pattern provides the PDC with some shipping flexibility, a costly overflow shipment of 4 units occurs on day 5.

An optimal solution from ($\mathbf{P}_\delta$) for Example 1 (starting from one of the randomly-generated initial sets of $\delta_{t'd}$ values) results in positive values of δ for nine of the 15 elements. For five of the nine positive δ values, the corresponding $x_{t'd}$ values are also positive. Thus, there are five relevant δ values, i.e., $\delta_{1,1} = 0.96$, $\delta_{2,2} = 1.00$, $\delta_{2,4} = 0.26$, $\delta_{3,3} = 1.00$, and $\delta_{4,4} = 1.00$, which correspond to net prices of \$39.04, \$39, \$38.64, \$39, and \$39, respectively (see Table 4). The optimal production quantities and package flows are shown in the lower portion of Figure 2. Interestingly, the PDC gives discounts for same-day releases as long as these releases are not on day 5, the day with the overflow shipments. Correspondingly, the customer takes advantage of the low-priced same-day release option on days 1 through 4. The customer shifts his production schedule slightly (producing two more units on day 4 and two fewer units on day 5) in order to avoid releases on day 5. The new release schedule allows the PDC to modify his shipping schedule to entirely eliminate overflow shipments.

In Example 2, the initial price vector also reflects speed-of-service pricing. Production and transportation capacities are moderately variable and are given by the vectors (6, 19, 17, 8, 0) and (13, 10, 18, 13, 8),

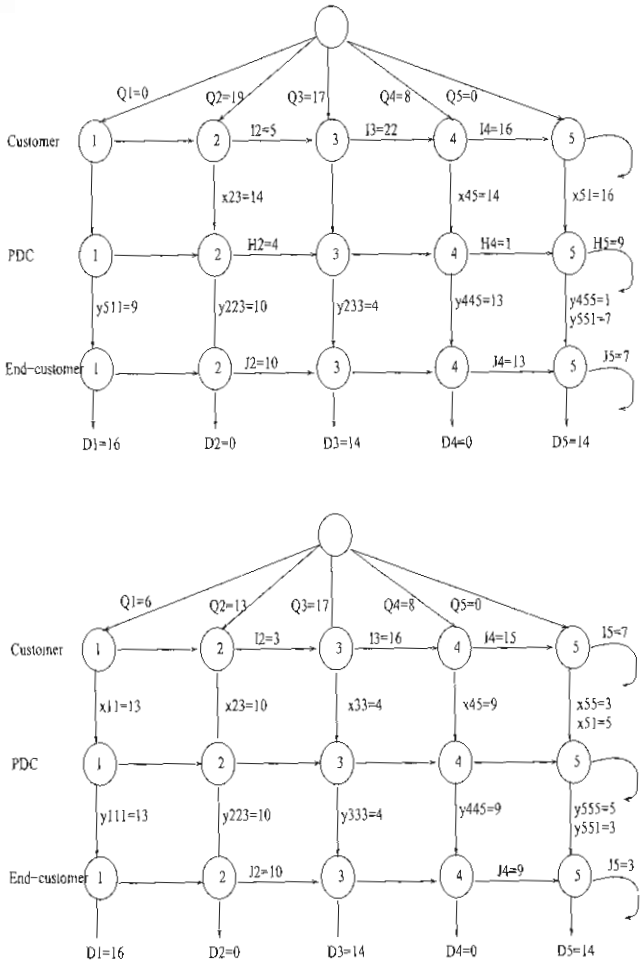
Figure 2 "Before" (upper) and "After" (lower) solutions for Example 1.



respectively. Demands are shown on the network in the upper portion of Figure 3, along with production quantities and product flows under $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. Under the stated speed-of-service pricing, the customer prefers to release packages one day before they are due, and the solution $\hat{\mathbf{x}}$ reflects this preference. The PDC is able to satisfy the customer's demands on time without using overflow shipments, but both parties incur inventory holding costs.

An optimal solution from ($\mathbf{P}_\delta$) for Example 2 (starting from one of the randomly-generated initial sets of $\delta_{t'd}$ values) results in positive values for δ for seven of the 15 elements. Of these seven, there are five (t, d) pairs for which the corresponding $x_{t'd}$ value is also positive. Thus, there are five relevant δ values, i.e., $\delta_{1,1} = 0.98$, $\delta_{3,3} = 1.00$, $\delta_{4,5} = 0.01$, $\delta_{5,1} = 0.10$, and $\delta_{5,5} = 0.99$, which correspond to net prices of \$39.02, \$39, \$38.99, \$38.90, and \$39.01, respectively (see Table 5). The optimal production quantities and package flows are shown in the lower portion of Figure 3. The customer changes his production schedule to better match his new package release schedule. The PDC

Figure 3 "Before" (upper) and "After" (lower) solutions for Example 2.



ships all packages as soon as they are received from the customer, thereby eliminating inventory storage costs. The customer benefits from reduced inventory holding costs as well as from a small discount on his total transportation cost. We found this example particularly interesting because the customer initially releases all packages one day early. Ordinarily, this would provide an acceptable level of flexibility for the PDC. However, in the improved solution, the customer is provided discounts for "just-in-time" releases because these are better aligned with the PDC's transportation capacity profile.

Table 4 Discounted Prices ($p_{t'd} - \delta_{t'd}$) for Example 1

Release day	Due date				
	Monday	Tuesday	Wednesday	Thursday	Friday
Monday	39.04	39.00	38.90	—	—
Tuesday	—	39.00	39.00	38.64	—
Wednesday	—	—	39.00	39.00	38.90
Thursday	38.90	—	—	39.00	39.00
Friday	39.00	38.90	—	—	40.00

Table 5 Discounted Prices ($p_{t'd} - \delta_{t'd}$) for Example 2

Release day	Due date				
	Monday	Tuesday	Wednesday	Thursday	Friday
Monday	39.02	39.00	38.90	—	—
Tuesday	—	40.00	39.00	38.90	—
Wednesday	—	—	39.00	39.00	38.90
Thursday	38.90	—	—	39.00	38.99
Friday	38.90	38.90	—	—	39.01

These two examples demonstrate that Pareto-improving solutions may have both intuitive and unintuitive characteristics. Common wisdom in the package delivery industry suggests that more flexibility and looser deadlines are advantageous. This turns out to be helpful in Example 1, where, although the customer releases some units later than they were released initially, other units are released earlier. The price discounts compensate the customer for adjusting his releases to provide the PDC more flexibility. On the other hand, contrary to common wisdom, in Example 2, the results suggest that it may even be advantageous for the PDC to give the customer incentives to alter his release pattern so that some releases are strictly later and none are earlier. More broadly, the results indicate that the PDC should neither aim for more flexibility nor have more "just-in-time" releases but should instead aim for a target with releases in "the right quantity at the right time," and that systematically-designed economic incentives to move the customer toward that target may have sizable benefits.

7. Discussion and Conclusions

We have developed an approach for an express package delivery company to use in structuring contracts for its large customers with stable demand patterns. The method utilizes day-of-week and speed-of-service pricing differentials as an element of the contract. The goal is to structure a contract that maximizes the package delivery company's profit while ensuring that the customer is as well off as he was under a reference price schedule (e.g., an existing contract). The customer is compensated for costly deviations of the new package release schedule from the initial schedule via price discounts that are determined within our solution approach.

Although the focus of this paper has been on improving operating profits for the PDC while increasing its competitive position with potential customers, there are side benefits of our methodology. By smoothing the demand for transport capacity, our approach may also reduce fixed capital and maintenance costs by limiting peak capacity requirements. This would allow the PDC to satisfy its demands with

a smaller fleet and/or with smaller vehicles on a given route.

We observed that for many problem instances, the discounts provided to the customer are zero or negligible. This implies that alternate optima or near-optima for $\mathbf{x}$ exist with $\delta = 0$ that are much less costly for the PDC. The results also suggest that in some cases, the potential savings for the PDC may be substantial, so in these instances it would be worthwhile for the PDC to offer the customer additional discounts, above and beyond the minimum required to ensure that the customer is as well off as he was under the original price schedule, as inducement to use the release schedule from (P_δ). Such an inducement may be necessary to prevent the customer from using his cost-minimizing release schedule under the discounted prices. Without voluntary compliance on the part of the customer to use $\mathbf{x}$ or a similar release schedule, the PDC would need to monitor the customer's releases in order to ensure that the full benefit of the contract terms can be achieved.

Alternate optima arise in a different form as well. Recall that we used 250 different starting points to solve each problem instance. Interestingly (although not surprisingly), we found many distinct alternate optima that have different δ values, different $\mathbf{x}$ vectors and different $\mathbf{y}$ vectors. We view the ability to generate multiple solutions as an important advantage of our approach: the package delivery company can offer the customer many different options that have equal or similar profit consequences for the PDC. This provides more flexibility in negotiations, and allows the two parties to evaluate the solutions on the basis of secondary criteria.

Our approach can be generalized in a straightforward manner to accommodate multiple types of packages. The resulting optimization problem would require more time to solve, but we expect the CPU times to be well within the limits considered acceptable for aiding contract negotiations that occur only once or twice a year for each customer. Because of our empirical findings that the single-customer, single-product price optimization problem is reasonably well-behaved, this provides hope for a generalization to multiple customers, multiple products, and multiple shared routes. As reported in the previous section, it is not difficult to find optimal or very near optimal prices despite the absence of any guarantees of unimodality and despite evidence that multiple optimal solutions may exist. Because quoted prices usually must be stated as a multiple of a basic monetary unit, and usually are rounded further to convenient values, we believe that carefully implemented search procedures may provide reasonably good solutions for realistic scenarios. However, further work is needed to obtain convincing evidence that the objective function is, indeed, well behaved in more complex settings.

Further research is needed to devise methods for selecting among alternate optima based on secondary criteria, such as smoothing the overall workload for the PDC or smoothing production for the customer, or selecting the pricing solution that is best suited for allowing the customer to independently make decisions while limiting the repercussions for the PDC.

Acknowledgments

We are grateful to the management and staff of an unnamed package delivery company for many useful discussions on this topic. We also acknowledge the help of Professor Nikolaos Sahinidis in implementing the BARON solver, and the thoughtful comments of two anonymous referees whose suggestions strengthened the paper.

References

Ahuja, R. K., J. B. Orlin. 2001. Inverse optimization. *Operations Research* 49(5) 771–783.

Al-Khayyal, F. A. 1990. Jointly constrained bilinear programs and related problems: An overview. *Computers and Mathematics with Applications* 19(11) 53–62.

Al-Khayyal, F. A., J. E. Falk. 1983. Jointly Constrained Biconvex Programming. *Mathematics of Operations Research* 8(2) 273–286.

Altmann, J., K. Chu. 2001. How to charge for network services—Flat-rate or usage-based? *Computer Networks* 36(5–6) 519–531.

Arnott, R., A. de Palma, R. Lindsey. 1990. Departure time and route choice for the morning commute. *Transportation Research B* 24(3) 209–228.

Boiteux, M. 1960. Peak-load pricing. *Journal of Business* 33(2) 157–179.

Borenstein, S. 2005. Time-varying retail electricity prices. Theory and practice; in *Electricity Deregulation*, J. M. Griffin, S. L. Puller (eds.), University of Chicago Press, Chicago, Illinois.

Bitran, G., R. Caldentey. 2003. An overview of pricing models for revenue management. *Manufacturing and Service Operations Management* 5(3) 203–229.

Brotcorne, L., M. Labbe, P. Marcotte, G. Savard. 2000. A bilevel model and solution algorithm for a freight tariff-setting problem. *Transportation Science* 34(3) 289–302.

Brotcorne, L., M. Labbe, P. Marcotte, G. Savard. 2001. A bilevel model for toll optimization on a multicommodity transportation network. *Transportation Science* 35(4) 345–58.

Button, K. J., E. T. Verhoef (eds.). 1998. *Road pricing, traffic congestion and the environment: Issues of efficiency and social feasibility*. Edward Elgar, Cheltenham, UK, and Northampton, Massachusetts.

Chen, H., M. Z. Frank. 2001. State dependent pricing with a queue. *IIE Transactions* 33(10) 847–860.

Chen, H., Y.-W. Wan. 2003. Price competition of make-to-order firms. *IIE Transactions* 35(9) 817–832.

Crew, M., C. Fernando, P. Kleindorfer. 1995. The theory of peak-load pricing: A survey. *Journal of Regulatory Economics* 8(3) 215–248.

Crew, M. A., P. R. Kleindorfer, M. A. Smith. 1990. Peak-load pricing in postal services. *The Economic Journal* 100(402) 793–808.

Daganzo, C., R. Garcia. 2000. A Pareto improving strategy for the time-dependent morning commute problem. *Transportation Science* 34(3) 303–311.

Eckel, C. 1987. Customer-class price discrimination by electric utilities. *Journal of Economics and Business* 39(1) 19–33.

Gavish, B., P. J. Schweitzer. 1977. The Markovian queue with bounded waiting time. *Management Science* 23(12) 1349–1357.

Gerstner, E. 1986. Peak load pricing in competitive markets. *Economic Inquiry* 24(2) 349–361.

Gupta, A., L. Linden, D. Stahl, A. Whinston. 2001. Benefits and costs of adopting usage-based pricing in a subnetwork. *Information Technology and Management* 2(2) 175–191.

Ha, A. Y. 2001. Optimal pricing that coordinates queues with customer-chosen service requirements. *Management Science* 47(7) 915–930.

Holguin-Verasa, J., S. Jara-Diaz. 1999. Optimal pricing for priority service and space allocation in container ports. *Transportation Research B* 33(2) 81–106.

Kamerschen, D. P., D. V. Porter. 2004. The demand for residential, industrial and total electricity, 1973–1998. *Energy Economics* 26(1) 87–100.

Li, L., Y. S. Lee. 1994. Pricing and delivery-time performance in a competitive environment. *Management Science* 40(5) 633–646.

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 Queue. *Operations Research* 38(5) 870–883.

Odlyzko, A. 2001. Internet pricing and the history of communications. *Computer Networks: The International Journal of Distributed Informatique* 36(5–6) 493–517.

Palaka, K., S. Erlebacher, D. H. Kropp. 1998. Lead-time setting, capacity utilization, and pricing decisions under lead-time dependent demand. *IIE Transactions* 30(2) 151–163.

Panzar, J. 1976. A neoclassical approach to peak-load pricing. *Bell Journal of Economics* 7(2) 521–530.

Paschalidis, I. C., J. Tsitsiklis. 2000. Congestion-dependent pricing of network services. *IEEE/ACM Transactions on Networking* 8(2) 171–184.

Plambeck, E. L. 2004. Optimal leadtime differentiation via diffusion approximations. *Operations Research* 52(2) 213–228.

Rao, S., E. R. Petersen. 1998. Optimal pricing of priority services. *Operations Research* 46(1) 46–56.

Sahinidis, N. V., M. Tawarmalani. 2005. BARON 7.2.5: Global optimization of mixed-integer nonlinear programs, *User's manual*. Available at http://www.gams.com/dd/docs/solvers/baron.pdf.

Stidham, Jr., S. 1985. Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control* 30(8) 705–713.

Stidham, Jr., S. 1992. Pricing and capacity decisions for a service facility: Stability and multiple local optima. *Management Science* 38(8) 1121–1139.

Talluri, K. T., G. J. van Ryzin. 2004. *The theory and practice of revenue management*. International series in operations research and management science, Vol. 68, Springer Science, New York, New York.

Vickrey, W. 1963. Pricing in urban and suburban transport. *American Economic Review* 53(2) 452–465.

Vickrey, W. 1971. Responsive pricing of public utility services. *The Bell Journal of Economics and Management Science* 2(1) 337–346.

Wachs, M. 2005. Then and now: The evolution of congestion pricing in transportation and where we stand today. Resource paper for the International Symposium on Road Pricing, Key Biscayne, Florida, November 2003; in *International Perspectives on Road Pricing Conference Proceedings* 34, Transportation Research Board, 63–72.

Weatherford, L. R., S. E. Bodily. 1992. A taxonomy and research overview of perishable-asset revenue management: Yield management, overbooking and pricing. *Operations Research* 40(5) 831–844.

Webster, S. 2002. Dynamic pricing and lead-time policies for make-to-order systems. *Decision Sciences* 33(4) 579–599.

Williamson, O. 1966. Peak load pricing and optimal capacity under indivisibility constraints. *American Economic Review* 56(4) 810–827.

Wilson, R. 1993. *Nonlinear pricing*, Oxford University Press, New York, New York, Chapters 9–11.

Yan, H., W. Lam. 1996. Optimal road tolls under conditions of queueing and congestion. *Transportation Research A* 30(5) 319–332.